

# P&S Mobile Genomics

## Lecture 11: Genomic Data Sharing Under Differential Privacy

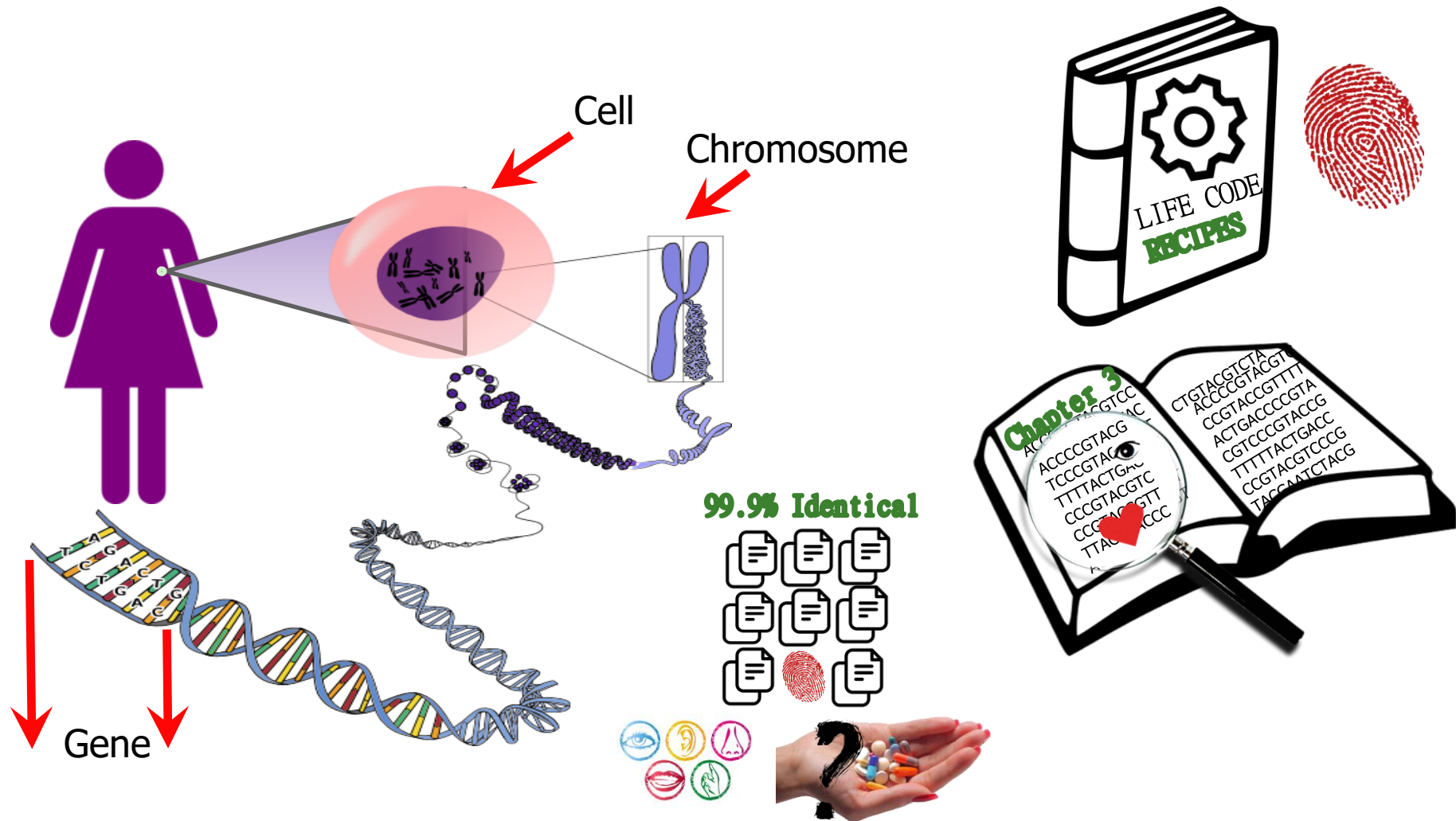
Dr. Nour Almadhoun Alserr

ETH Zurich

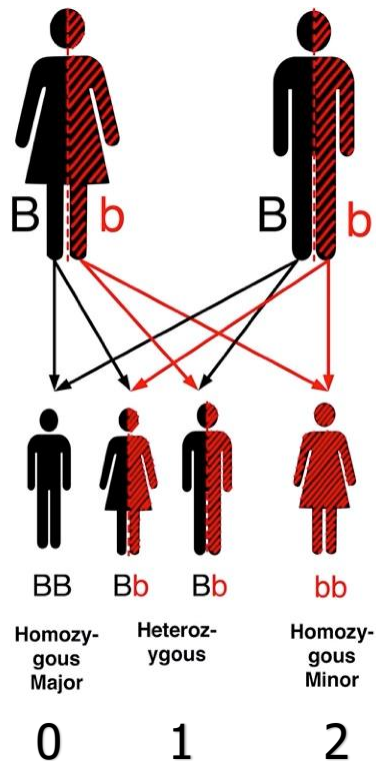
Fall 2022

10 January 2023

# Genome



# Mendel's Law

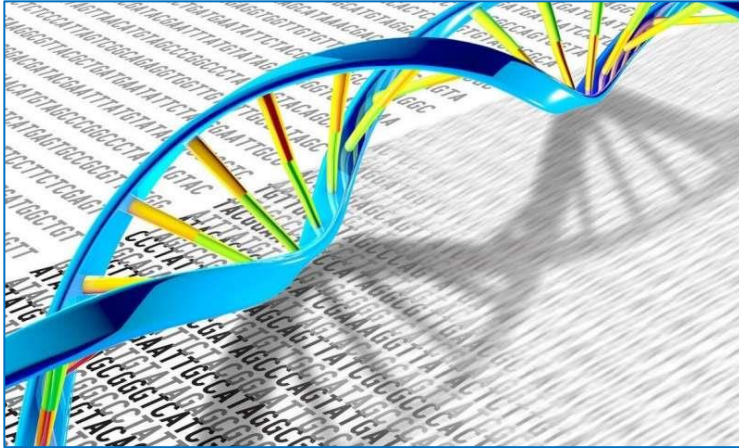


Mother	Father		
	BB	Bb	bb
BB	(1,0,0)	(0.5,0.5,0)	(0,1,0)
Bb	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
bb	(0,1,0)	(0,0.5,0.5)	(0,0,1)

Mother	Child		
	BB	Bb	bb
BB	(0.5,0.5,0)	(0,0.5,0.5)	N/A
Bb	(0.5,0.5,0)	(0.33,0.33,0.33)	(0,0.5,0.5)
bb	N/A	(0.5,0.5,0)	(0,0.5,0.5)

# The Genomic Era

---



© Medical Press

2025

1 Zetta-Bases/year ( $10^{21}$ ) capacity  
105 Million Sequenced Human genome

# The Genomic Era



[Home](#) > [Genomics](#) > 100,000 Genomes Project

## 100,000 Genomes Project

The 100,000 Genomes Project is cementing the NHS's position as one of the most advanced healthcare systems in the world, and is providing the foundation for a new era of [personalised medicine](#), and this in turn will contribute towards delivering high quality care for all, now and for future generations.

The 100,000 Genomes Project aims to bring the benefits of personalised medicine to the NHS. To make sure patients benefit from innovations in genomics, the Government has committed to sequencing 100,000 whole human genomes, from 70,000 patients, by the end of 2018.

## European '1+ Million Genomes' Initiative

The Signatories of the declaration of cooperation "Towards access to at least 1 million sequenced genomes in the EU by 2022" are setting up a collaboration mechanism with the potential to improve disease prevention, allow for more personalised treatments and provide a sufficient scale for new clinically impactful research.

69–92% of the respondents in these studies had positive attitudes towards genomics research and donating their DNA samples.

---

## **A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States**

Nanibaa' A. Garrison, PhD<sup>1,2</sup>, Nila A. Sathe, MA, MLIS<sup>3,4</sup>, Armand H. Matheny, Antonmaria, MD, PhD<sup>5</sup>,  
Ingrid A. Holm, M  
Melissa L.

## **Genetic research participation in a young adult community sample**

Carla L. Storr • Flora Or • William W. Eaton •  
Nicholas Ialongo

## **Relationships in Medicine**

Meghan Halley, Nina  
ilfond & Sandra Soo-

nie, Meghan Halley, Nina

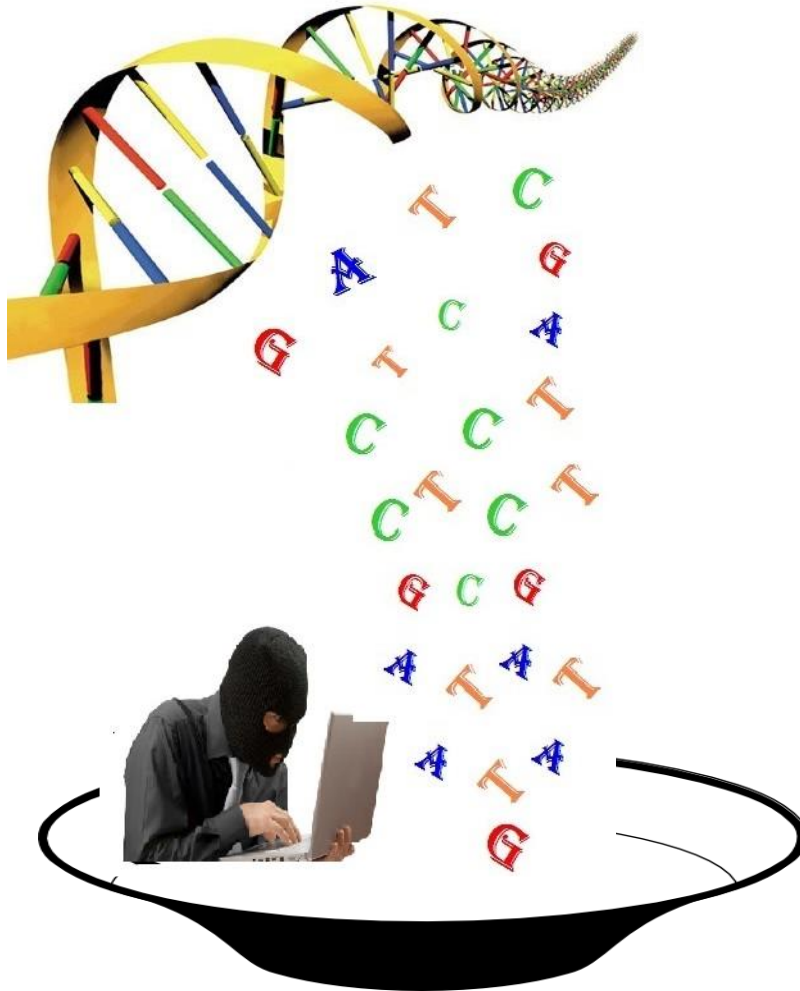
## **The Geisinger an electronic health record–linked biobank for precision medicine research**

David J. Carey, PhD<sup>1</sup>, Samantha N. Fetterolf, BS<sup>1</sup>, F. Daniel Davis, PhD<sup>1</sup>, William A. Faucett, MS<sup>1</sup>,  
H. Lester Kirchner, PhD<sup>1</sup>, Uyenlinh Mirshahi, PhD<sup>1</sup>, Michael F. Murray, MD<sup>1</sup>, Diane T. Smelser, PhD<sup>1</sup>,  
Glenn S. Gerhard, MD<sup>2</sup> and David H. Ledbetter, PhD<sup>1</sup>

## **f patient population**

Asa Kettils-Lindblad, Lena Ring, Eva Viderth, Mats G. Hansson

# Privacy Risks

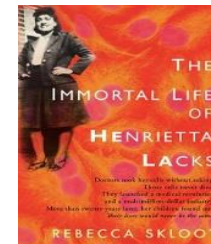


If it's on the Internet, it isn't private.



## If the owner of a genome is identified:

- He/she will face the risk of discrimination by employers or insurance companies.
- DNA sequences are highly correlated to the relatives' sequences, so relative's privacy will be at risk (Henrietta Lacks).





# Genome-Wide Association Study (GWAS)

Detecting genetic variants associated with phenotypes using two groups of people.

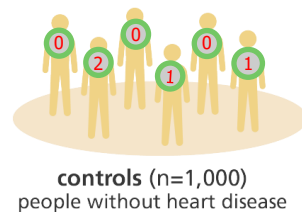
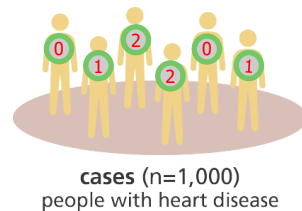
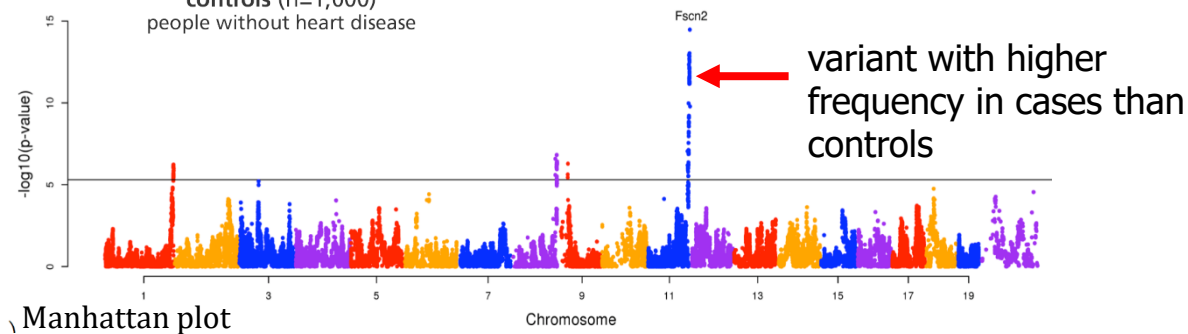


Table 1. GWAS genotype distribution for a  $2 \times 3$  contingency table (left) and a  $2 \times 2$  contingency table (right).

	Genotype			Total
	0	1	2	
Case	$S_0$	$S_1$	$S_2$	$S$
Control	$C_0$	$C_1$	$C_2$	$C$
Total	$n_0$	$n_1$	$n_2$	$n$

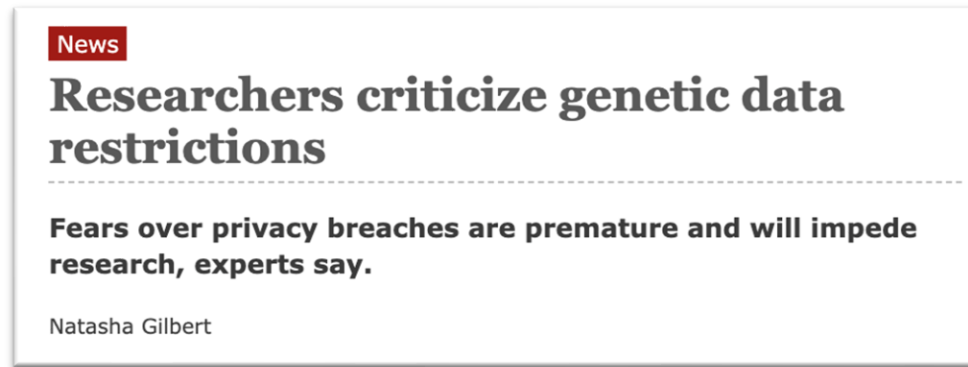
	Genotype		Total
	0	1	
Case	$S_0$	$S_1 + S_2$	$S$
Control	$C_0$	$C_1 + C_2$	$C$
Total	$n_0$	$n_1 + n_2$	$n$





# Genetic Data Restriction

---



- Researchers have assumed that case-control studies are safe to publish aggregate statistics of SNPs. Such belief was challenged when **Homer Attack** happened.
- **NIH** restricts the access to key results and data of GWAS to only trusted individuals.

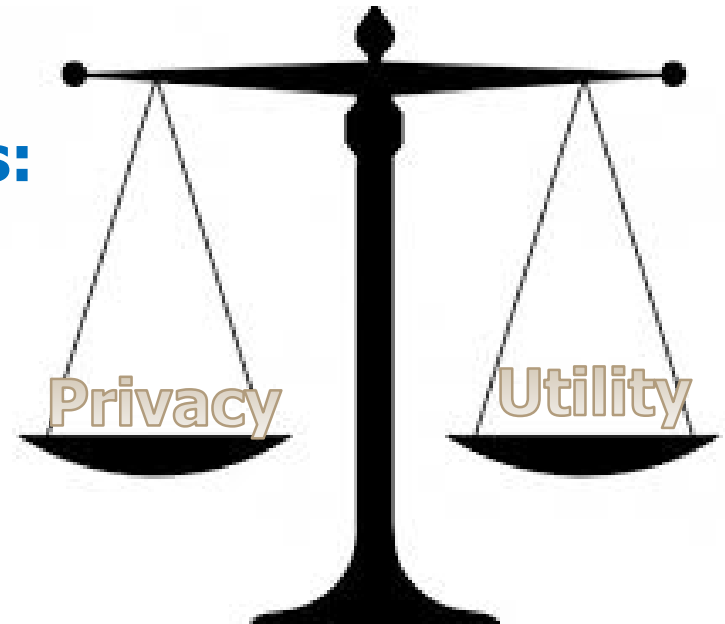
# Privacy-Utility Tradeoff

---

- Hiding some important data needs to tradeoff between **privacy** and **utility**.

➤ **Privacy preserving techniques:**

- K-anonymity.
- l-diversity.
- t-closeness.
- **Differential privacy.**
- Crypto-based techniques.

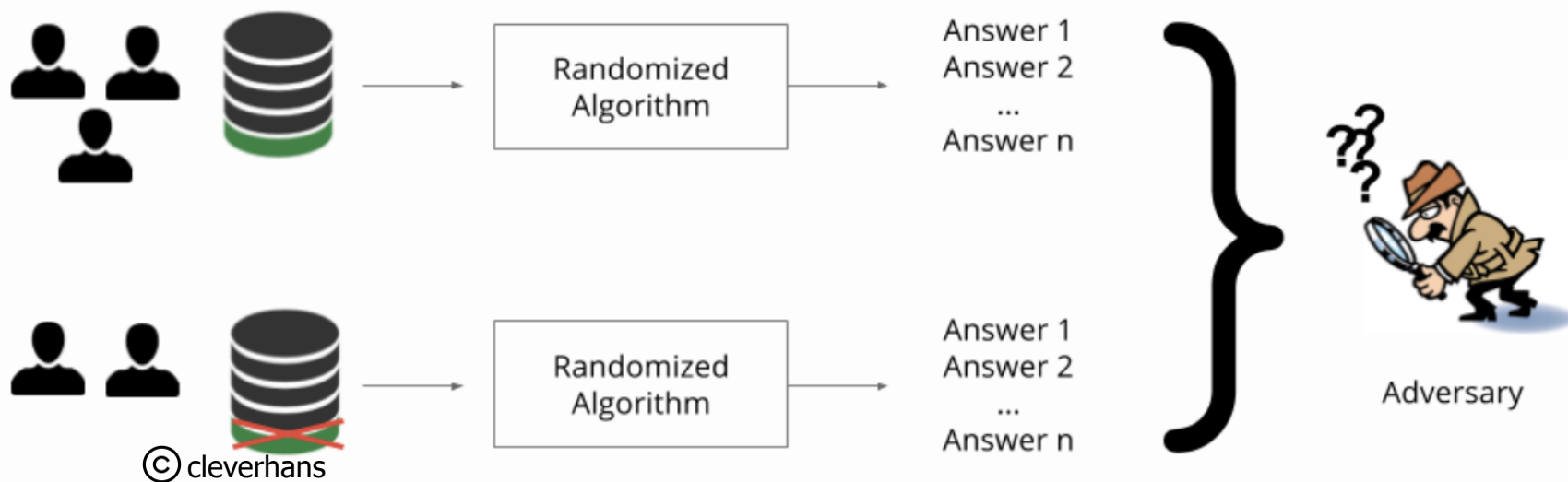




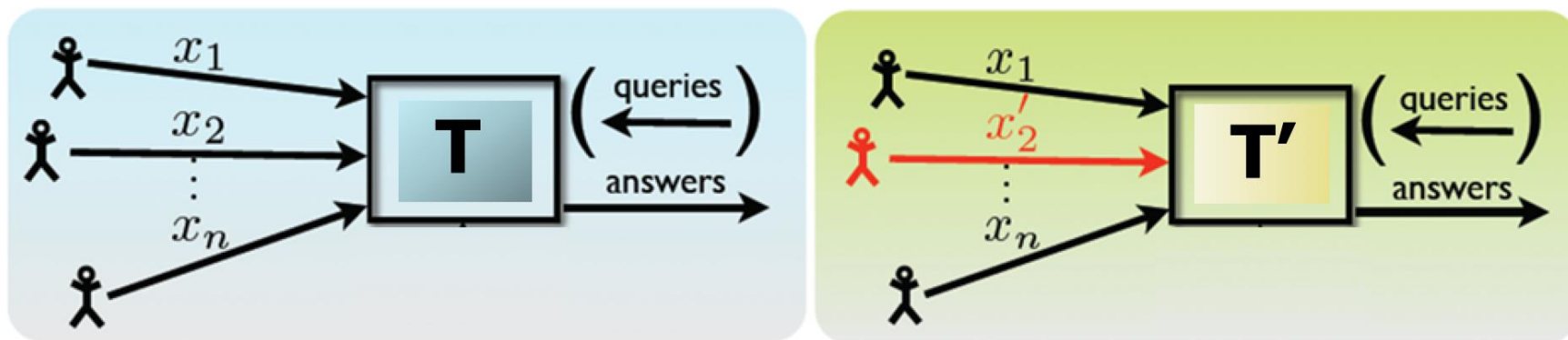
**Privacy-Preserving**

**Computing**  **Sharing**

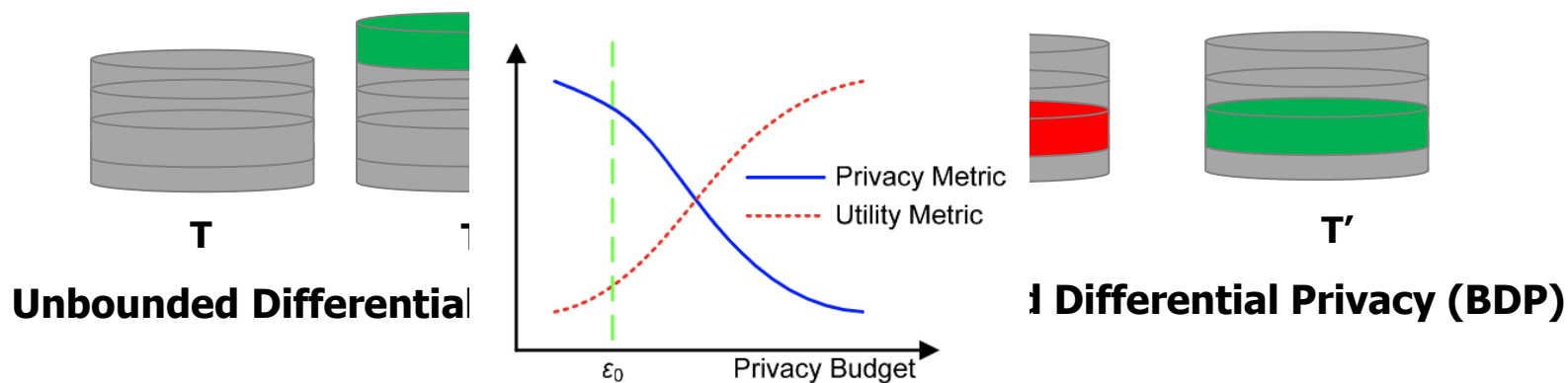
# Differential Privacy



# Differential Privacy



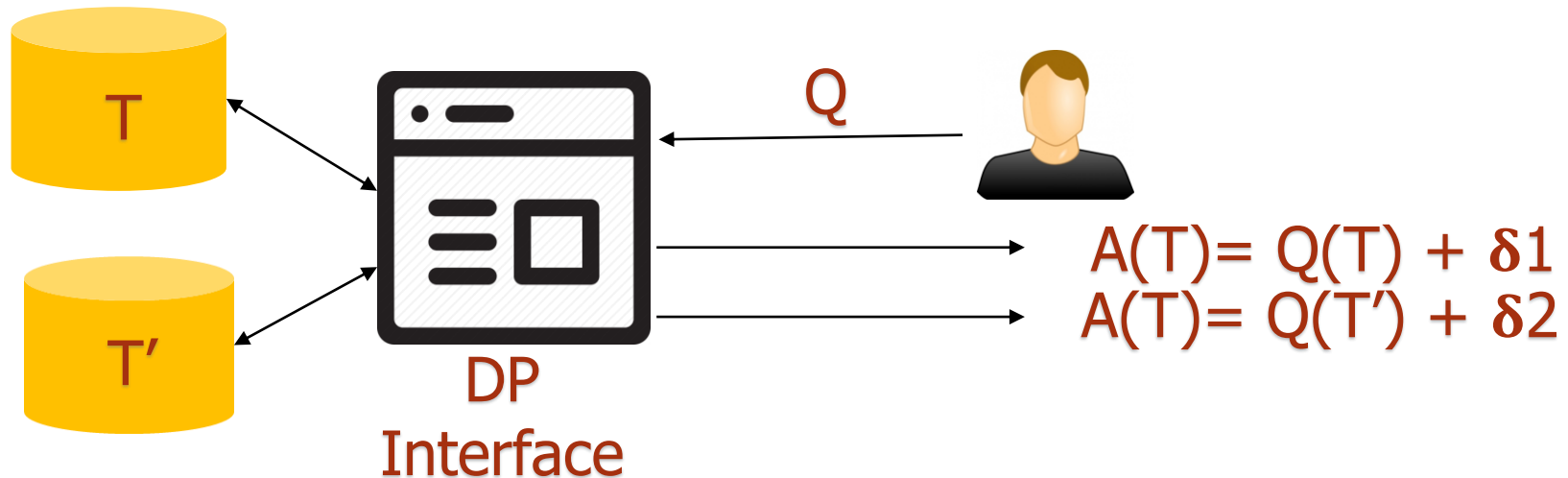
$$\Pr[A(T) \in O] \leq e^\epsilon \Pr[A(T') \in O]$$



© Liu, et al. *Applied Sciences* (2018)

# Laplace Perturbation Mechanism (LPM)

- $Q(T) + \delta$  where  $\delta$  is drawn from a Laplace distribution with mean 0 and scale  $\Delta Q/\epsilon$
- $\Delta Q$  : query global sensitivity



# Differential Privacy

---



Differential privacy library

Google Developer (2019)



Differential Privacy Team,  
Apple (2017)



Collecting Telemetry Data Privately (2017)



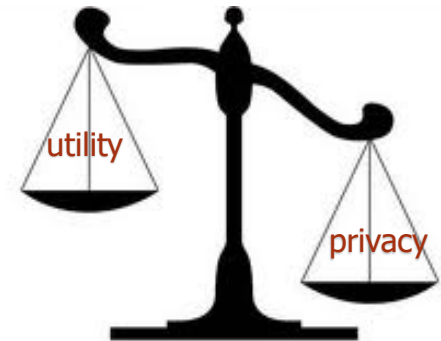
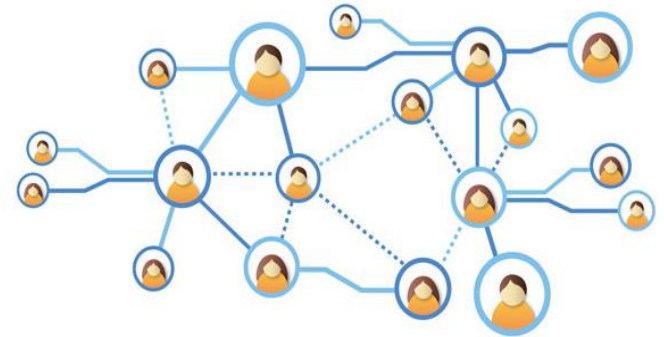
Differentially Private Publication System (2018)



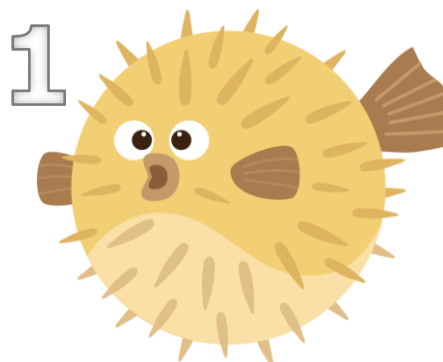
# Research Problem

---

- DP standard mechanism does not consider the **dependency between the data tuples** in the dataset.
- Current DP-based mechanisms which consider the tuples correlation, provide **poor accuracy**.



# Related Works



A generalization of DP



No perturbation algorithm

## **Pufferfish Framework**

(Kifer and Machanavajjhala, 2012)



Perturbation mechanisms



Deterministic constraints for the adversary

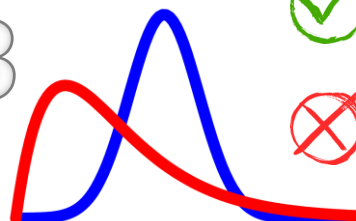
2



## **Blowfish Framework**

(He et al., 2014)

3



Perturbation mechanisms

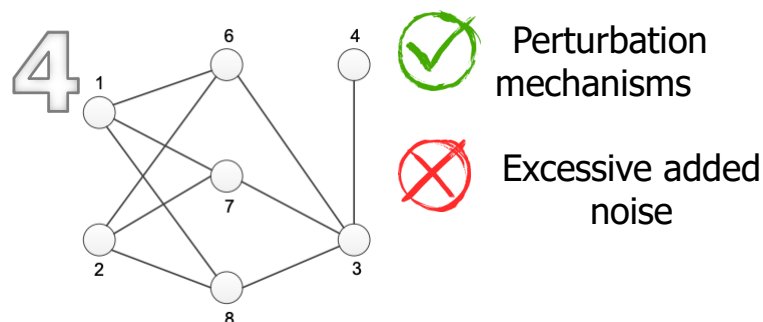


Correlation modeled by Gaussian Markov Random Fields

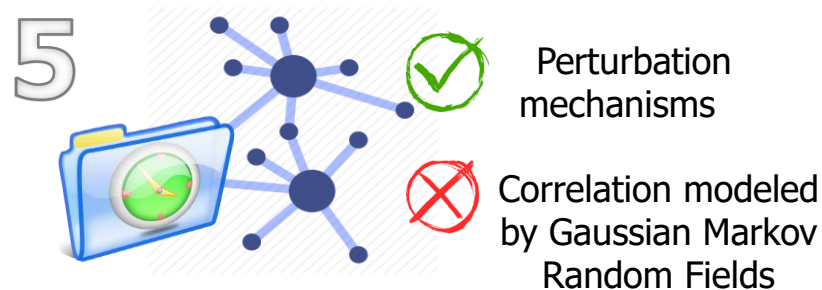
## **Bayesian DP**

(Yang et al., 2015)

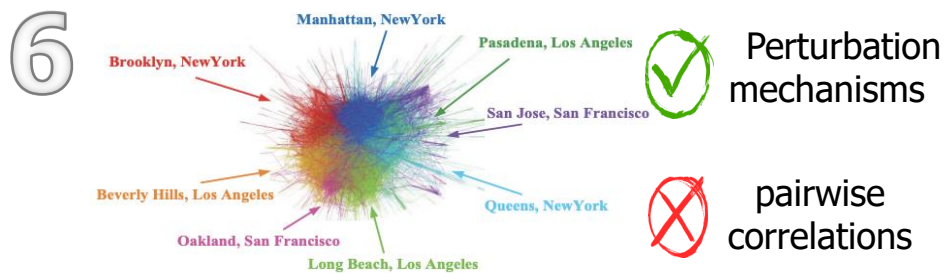
# Related Works



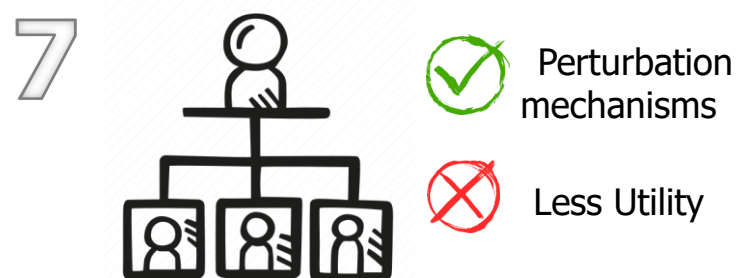
**Network Correlation**  
(Chen et al., 2014)



**Temporal Correlation**  
(Cao et al., 2017)



**Dependent DP**  
(Liu et al., 2016)



**Dependent DP**  
(Zhao et al., 2017)

# Our Contributions

---

## Attribute Inference Attack

**1** Differentially private **SUM** query results in a static genomic dataset with dependent tuples.

[Bioinformatics'19]

**2** Differentially private **MAF** and  $\chi^2$  query results in a static genomic dataset with dependent tuples.

[Bioinformatics'20] [ISMB'20]



## Membership Inference Attack

**3** Differentially private **MAF** in a static genomic dataset.

[Bioinformatics'20]

[ISMB'20]

# Our Contributions

---

## Countermeasures



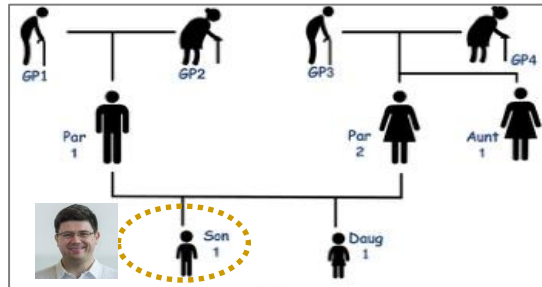
**4**  $\epsilon$ -differential privacy for sharing genomic datasets with dependent tuples .

**[Bioinformatics'19]**

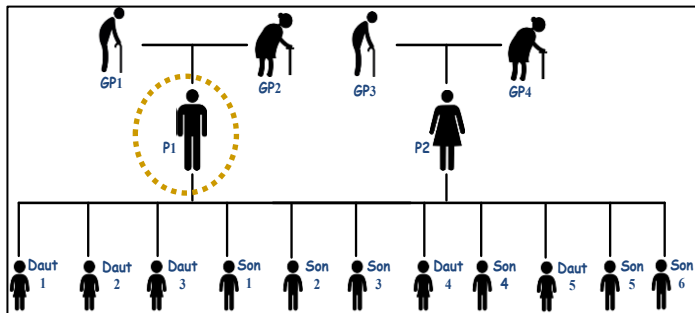
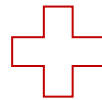
**5** Selective hiding mechanism and differential privacy.

**[arXiv'21]**

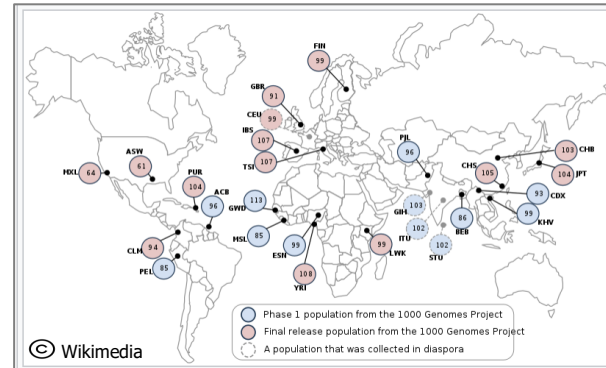
# Dataset Description



**Manuel Corpas Family**

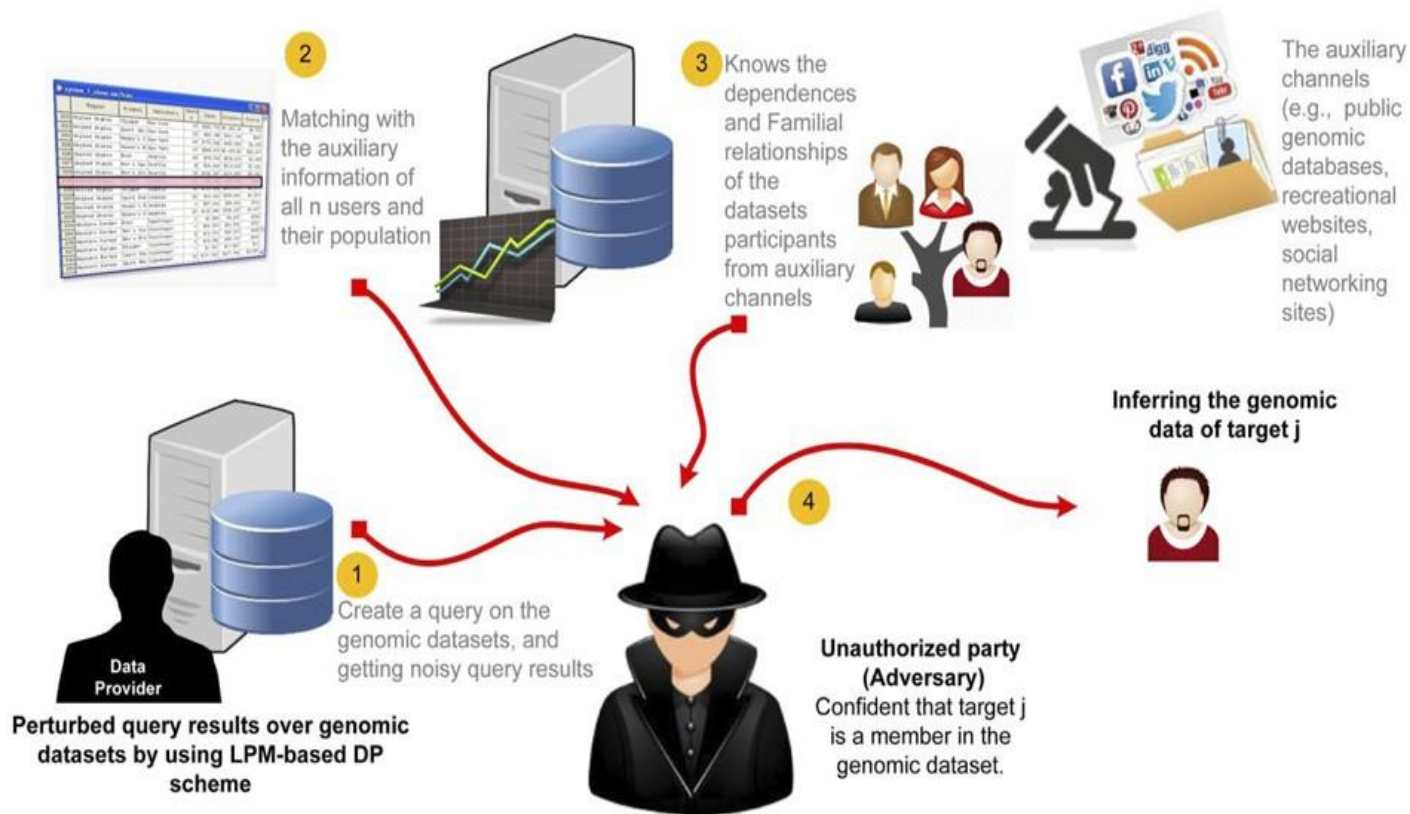


**CEPH/Utah Family**



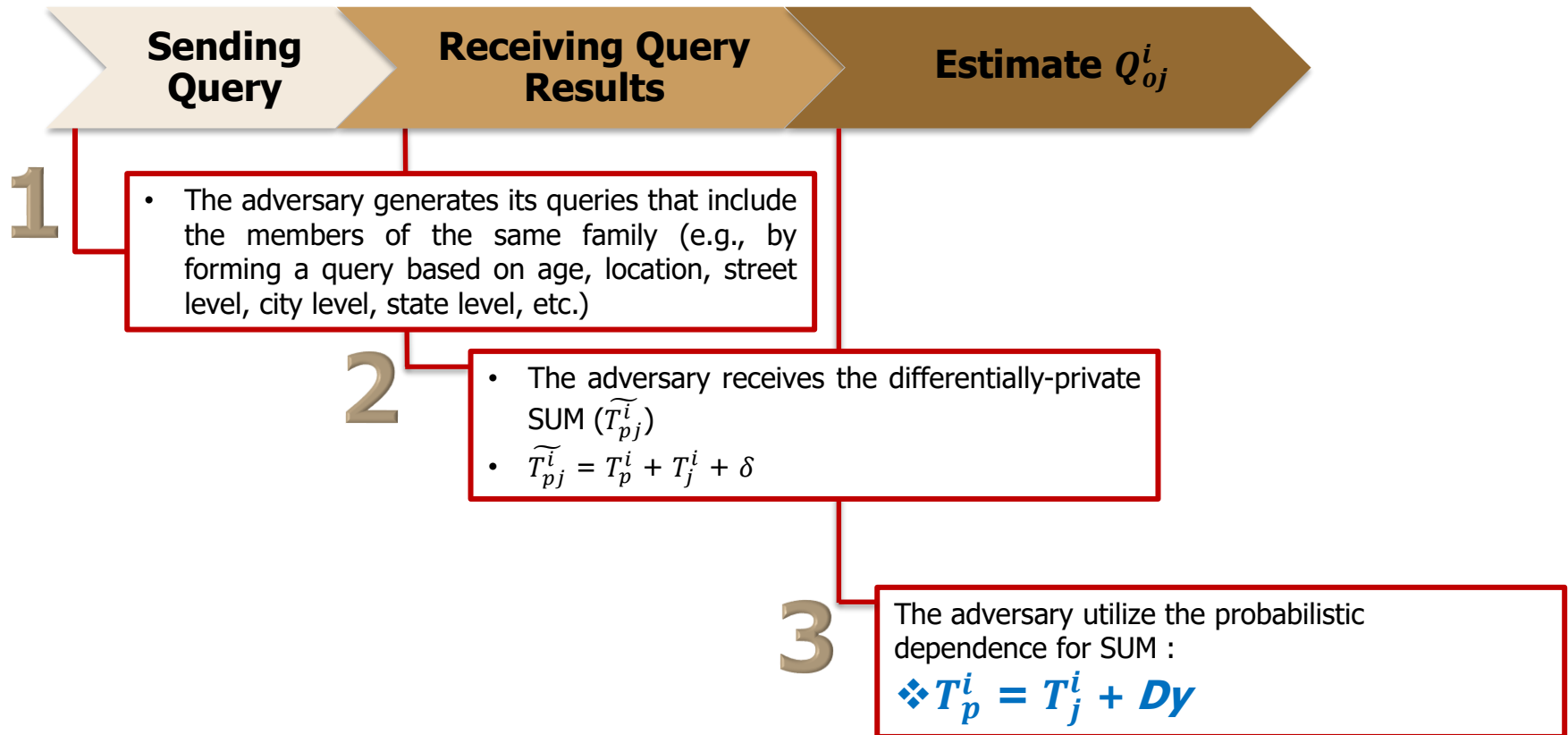
**n= 2514 or  
n= 2508,  
m =100**

# DP Inference Attacks





# Attribute Inference Attack



# Attribute Inference Attack

**Coin Change**

**Check Validity**

**Quantify the Attack Success**

4

- The adversary obtains all the possible partitions of  $T_{pj}^i$  (each partition will include  $(p+1)$  individuals).

$T_{pj}^i$ (Sum)	p+1 participants				
✓ 6	4				
✓ 6	4				
✗ 6	4	1		3	1

- The adversary uses Mendel's law to find the valid permutations for each partition. Then, he computes the probability by considering potential values of SNP  $i$  (0, 1, 2) for target  $j$ .

Father	Mother	Son	Son
✓ 1	2	2	1
✗ 1	2	1	0

- Estimation error metric:

$$E = \sum_{j=1}^m P(x_j^i | X_j) |Dist(x_j^i, x_j'^i)|$$

- Leaked information metric

$$L = \sum_{i=1}^m 1 - |sgn(Dist(x_j^i, x_j'^i))|$$

# Key Results

---

The adversary can infer the actual value of the targeted SNPs by up to **50%**.

Our proposed mechanism can achieve up to **50%** better privacy guarantees than the traditional DP-based solutions.

# DP Inference Attacks

---

**Nour Almadhoun**, Erman Ayday, and Ozgur Ulusoy

**[“Differential privacy under dependent tuples—the case of genomic privacy”](#)**

Bioinformatics, 2020

[[Source code](#)]

Bioinformatics



**Differential privacy under dependent tuples—  
the case of genomic privacy** FREE

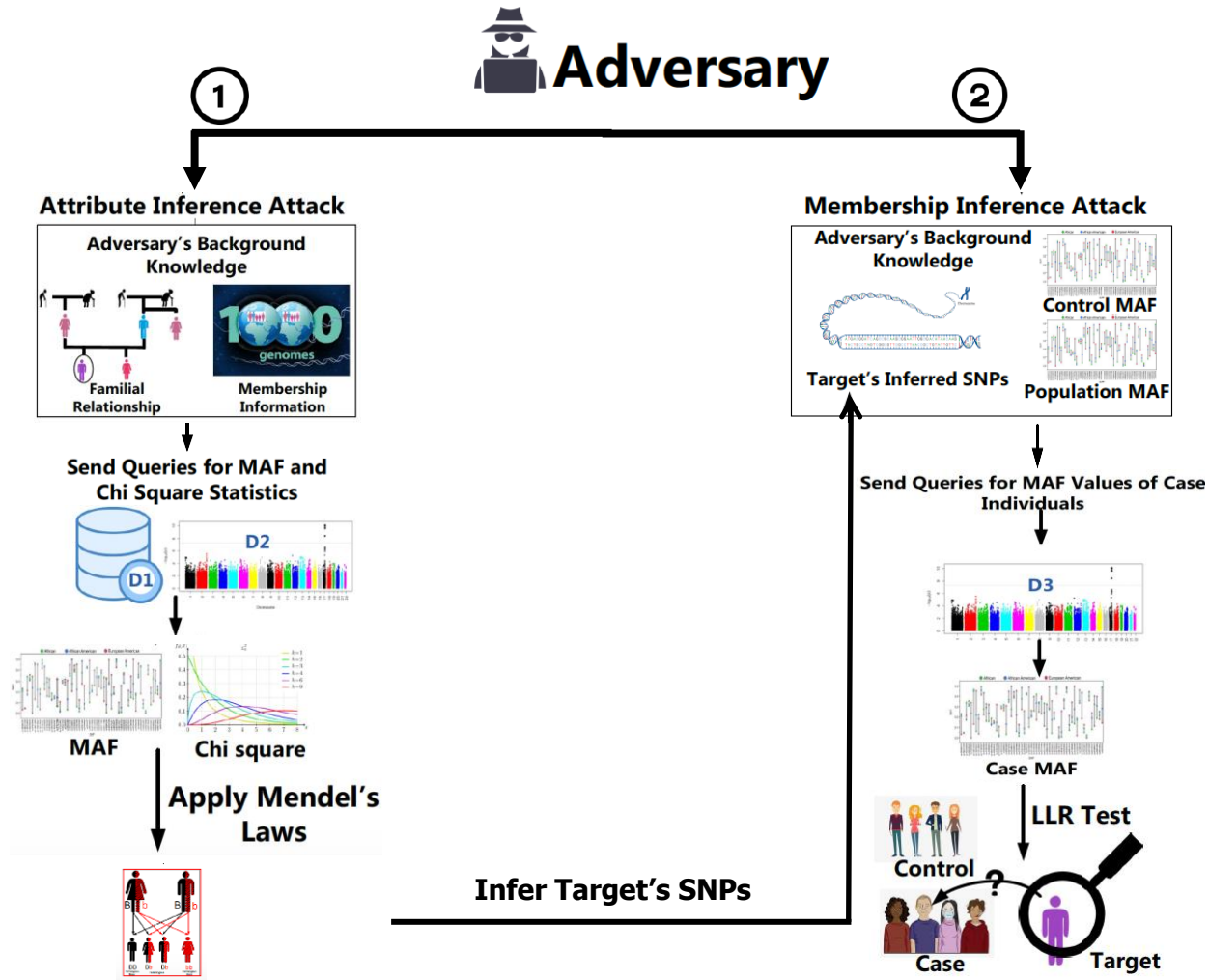
Nour Almadhoun, Erman Ayday ✉, Özgür Ulusoy ✉

*Bioinformatics*, Volume 36, Issue 6, 15 March 2020, Pages 1696–1703,

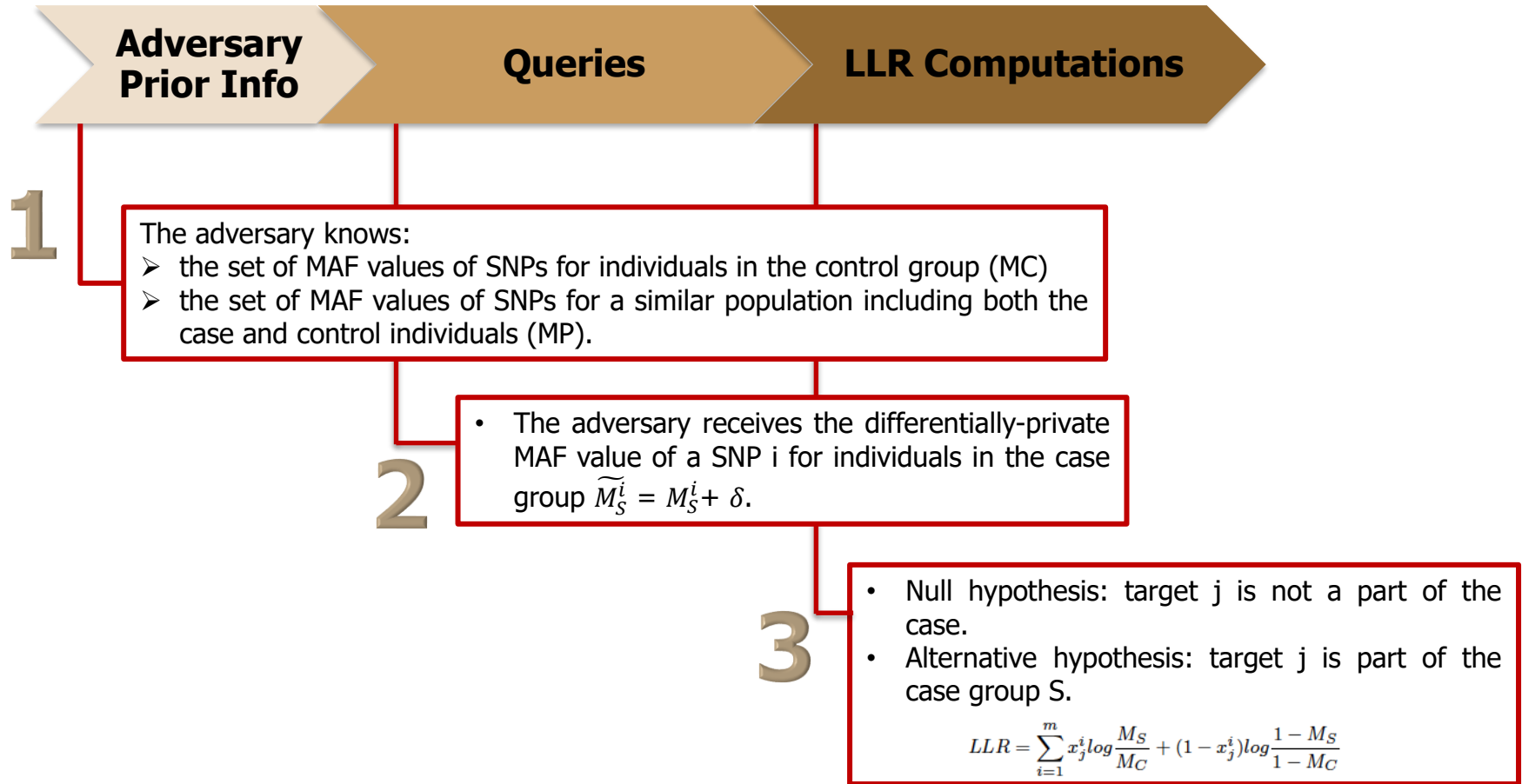
<https://doi.org/10.1093/bioinformatics/btz837>

**Published:** 08 November 2019    **Article history** ▼

# Threat Model



# Membership Inference Attack



# Key Results

---

An adversary can reveal up to **40% ~ 50%** more sensitive information about the genome of a target (compared to original privacy guarantees of standard DP-based mechanisms).

The inference power of the adversary can be **significantly high** in the membership attack even using inferred (and hence partially incorrect) genomes.



# DP Inference Attacks

---

**Nour Almadhoun**, Erman Ayday, and Ozgur Ulusoy

**["Inference attacks against differentially private query results from genomic datasets including dependent tuples"](#)**

Bioinformatics, 2020

[[Source code](#)]

Bioinformatics



**Inference attacks against differentially private query results from genomic datasets including dependent tuples** 

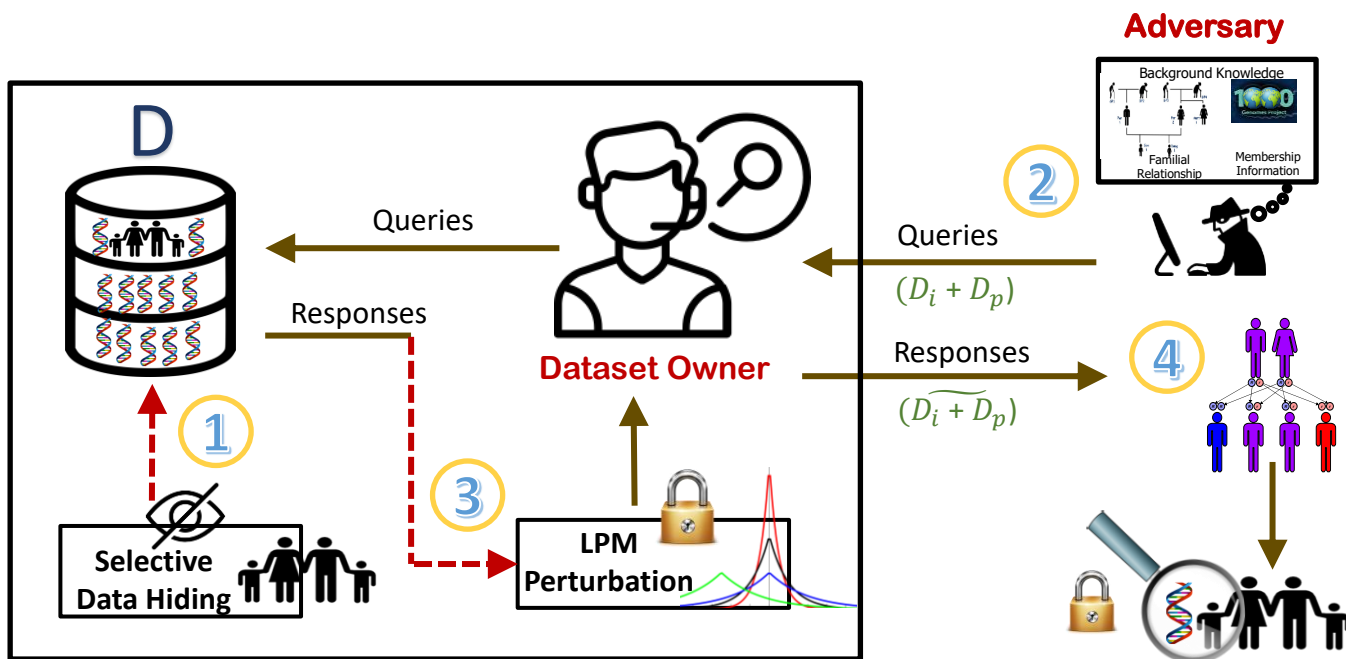
Nour Almadhoun, Erman Ayday , Özgür Ulusoy 

*Bioinformatics*, Volume 36, Issue Supplement\_1, July 2020, Pages i136–i145,

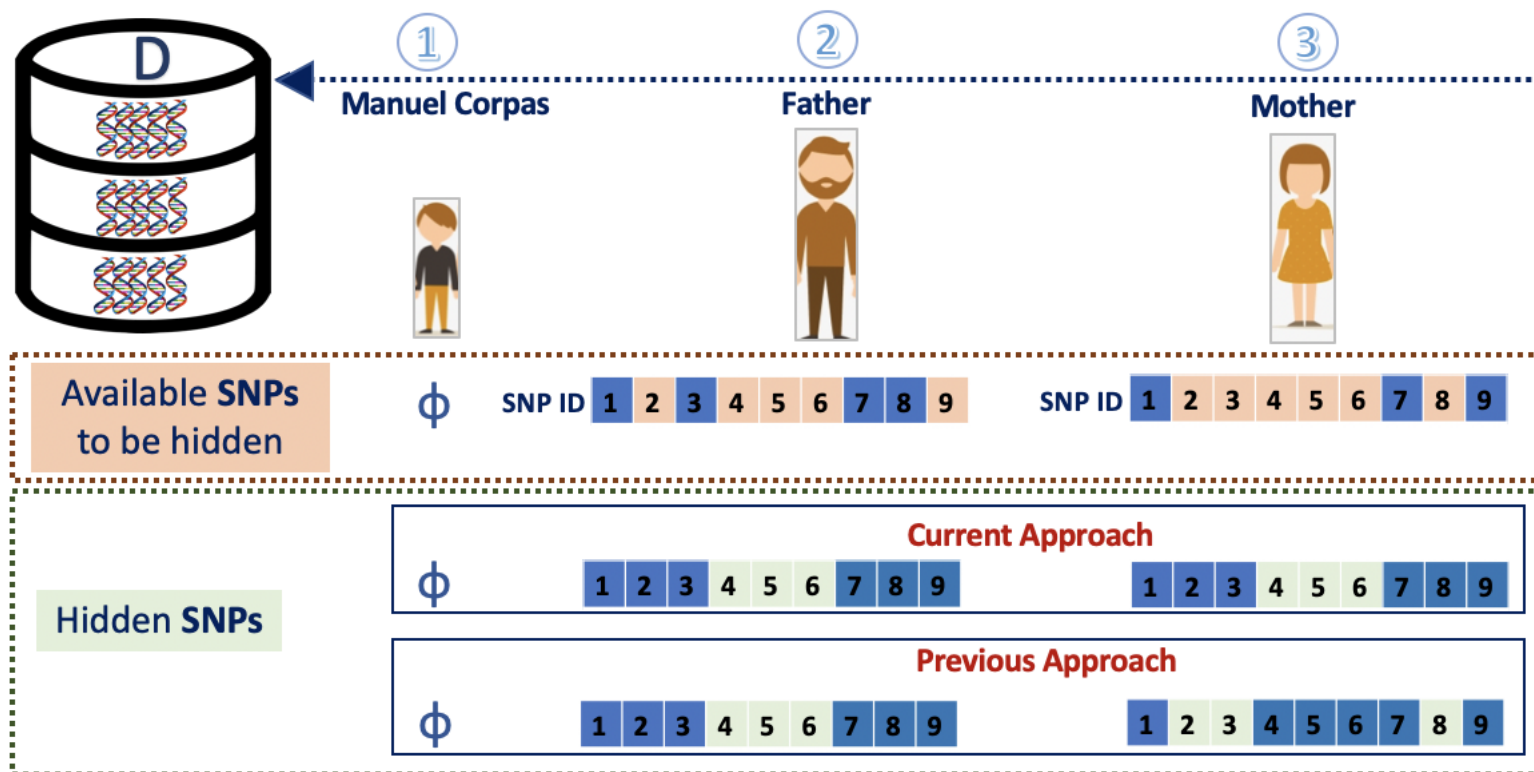
<https://doi.org/10.1093/bioinformatics/btaa475>

**Published:** 13 July 2020

# Selective Hiding Model



# Selective Hiding Model



# Key Results

---

We provide **similar privacy guarantees** of  $\epsilon$ -differential privacy, with **higher utility** than the state-of-the-art schemes.

# Selective SNP Hiding

---

**Nour Almadhoun Alserr**, Gulce Kale, Onur Mutlu, Oznur Tastan, Erman Ayday  
**[“Near-Optimal Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding”](#)**  
arXiv, 2021  
**[\[Source code\]](#)**

arXiv.org > cs > arXiv:2106.05211

Computer Science > Cryptography and Security

*[Submitted on 9 Jun 2021]*

**Near-Optimal Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding**

Nour Almadhoun Alserr, Gulce Kale, Onur Mutlu, Oznur Tastan, Erman Ayday

# GenShare Model

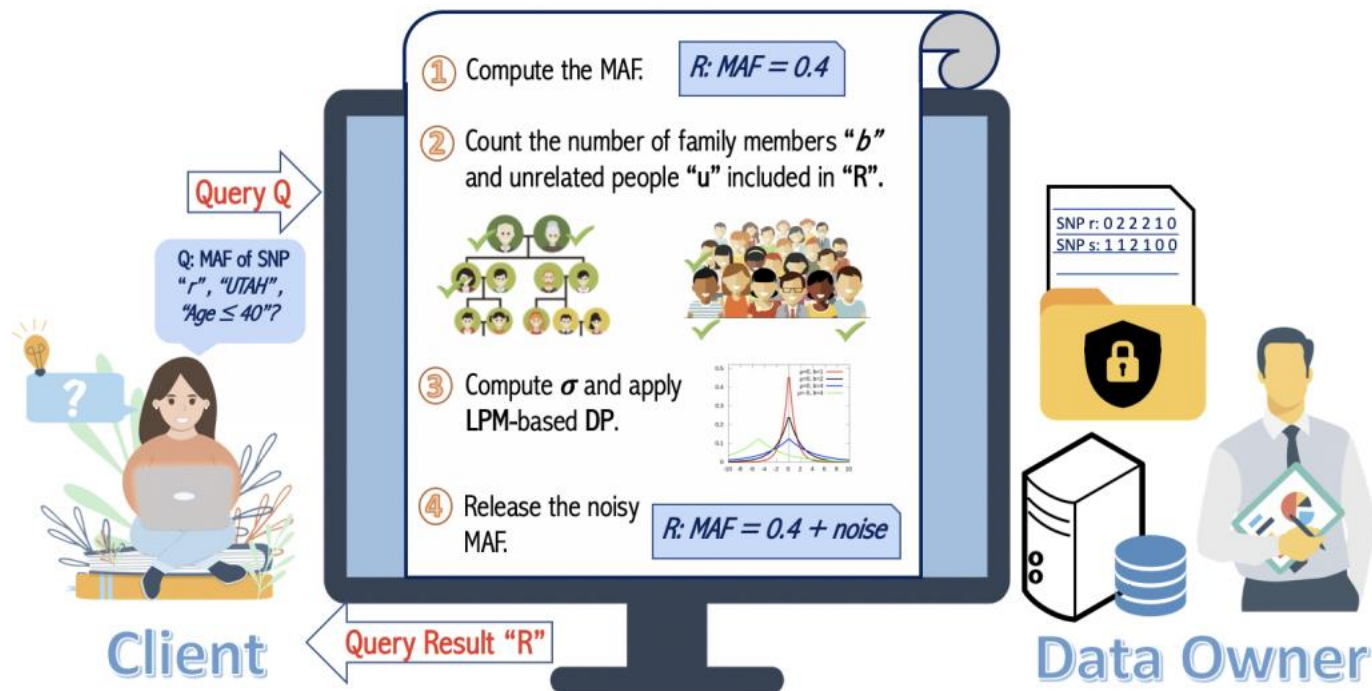


Fig. 1: Our proposed GenShare model

**Nour Almadhoun Alserr, Ozgur Ulusoy, Erman Ayday, Onur Mutlu**

**[“GenShare: Sharing Accurate Differentially-Private Statistics for Genomic Datasets with Dependent Tuples”](#)**

arXiv, 2021

arXiv > q-bio > arXiv:2112.15109

Quantitative Biology > Genomics

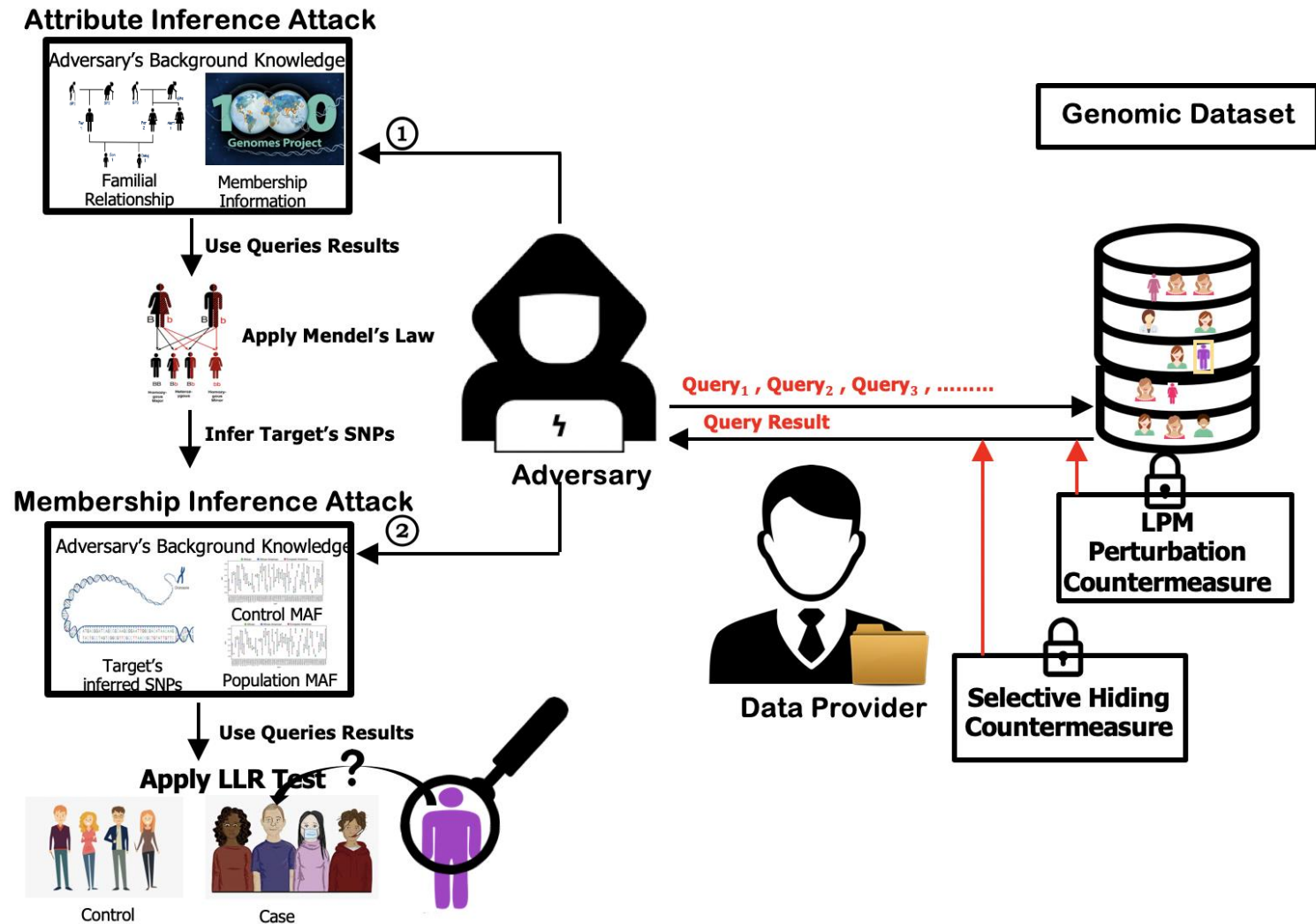
*[Submitted on 30 Dec 2021]*

**GenShare: Sharing Accurate Differentially-Private Statistics for Genomic Datasets with Dependent Tuples**

Nour Almadhoun Alserr, Ozgur Ulusoy, Erman Ayday, Onur Mutlu



# Full Model



# P&S Mobile Genomics

## Lecture 11: Genomic Data Sharing Under Differential Privacy

Dr. Nour Almadhoun Alserr

ETH Zurich

Fall 2022

10 January 2022